Chapter_8_HW

Christophe

November 24, 2015

8.2 Baby weights, Part II.

Exercise 8.1 introduces a data set on birth weight of babies. Another variable we consider is parity, which is 0 if the child is the first born, and 1 otherwise. The summary table below shows the results of a linear regression model for predicting the average birth weight of babies, measured in ounces, from parity.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	120.07	0.60	199.94	0.0000
parity	-1.93	1.19	-1.62	0.1052

(a) Write the equation of the regression line.

$\hat{birth\ weight} = 120.07 - 1.93 * parity$

(b) Interpret the slope in this context, and calculate the predicted birth weight of first borns and others.

> The estimated body weight of babies born that are the first born ("parity") is 1.93 ounces lower than babies born that are not first born.

First born: $120.07 - 1.93 * 1 = 118.14$

Not first born: $120.07 - 1.93 * 0 = 120.07$

(c) Is there a statistically significant relationship between the average birth weight and parity?

$H_O : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

$T = -1.62$

> The p value is .1052, because the p value is so large we fail to reject the null hypthosis. The data provided is not strong evidence that the true slope parameter is different than 0 and there is not an association between birth weight and parity.

8.4 Absenteeism, Part I.

Researchers interested in the relationship between absenteeism from school and certain demographic characteristics of children collected data from 146 randomly sampled students in rural New South Wales, Australia, in a particular school year. Below are three observations from this data set.

	eth	sex	lrn	days
1	0	1	1	2
2	0	1	1	11
⋮	⋮	⋮	⋮	⋮
146	1	0	0	37

The summary table below shows the results of a linear regression model for predicting the average number of days absent based on ethnic background (eth: 0 - aboriginal, 1 - not aboriginal), sex (sex: 0 - female, 1 - male), and learner status (lrn: 0 - average learner, 1 - slow learner).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18.93	2.57	7.37	0.0000
eth	-9.11	2.60	-3.51	0.0000
sex	3.10	2.64	1.18	0.2411
lrn	2.15	2.65	0.81	0.4177

(a) Write the equation of the regression line.

$\hat{absenteeism} = 18.93 - 9.11 * eth + 3.1 * sex + 2.15 * lrn$

(b) Interpret each one of the slopes in this context.

The estimated days of absenteeism of for a non-aboriginal is 9 days less than an aboriginal. Also, males are likely to be 3 days more absent than a female and slow learners are likely to be absent 2 days more than an average learner.

(c) Calculate the residual for the first observation in the data set: a student who is aboriginal, male, a slow learner, and missed 2 days of school.

$\hat{absenteeism} = 15.07$

$absenteeism_{case\ 1} = 2$

$e = -13.07$

The model overpredicts absenteeism for this case.

(d) The variance of the residuals is 240.57, and the variance of the number of absent days for all students in the data set is 264.17. Calculate the R2 and the adjusted R2. Note that there are 146 observations in the data set.

$R^2 = 1 - \frac{variability\ in\ residuals}{variability\ in\ the\ outcome}$

$R^2 = 0.0893$

$R^2_{adj} = 0.070097$

8.8 Absenteeism, Part II. Exercise 8.4 considers a model that predicts the number of days absent using three predictors: ethnic background (eth), gender (sex), and learner status (lrn). The table below shows the

adjusted R-squared for the model as well as adjusted R-squared values for all models we evaluate in the first step of the backwards elimination process.

	Model	Adjusted R^2
1	Full model	0.0701
2	No ethnicity	-0.0033
3	No sex	0.0676
4	No learner status	0.0723

Which, if any, variable should be removed from the model first?

Remove `No learner status` because it has a greater adjusted R^2 than the full model.

8.16 Challenger disaster, Part I.

On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. Temp gives the temperature in Fahrenheit, Damaged represents the number of damaged O-rings, and Undamaged represents the number of O-rings that were not damaged.

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

(a) Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

It appears that damaged o-rings occur more often when the tempurature drops below 66 degrees. At 53 degrees o-rings will become damanged very often. The damage above 66 degrees doesnt appear to be correlated with temperature.

(b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below. Describe the key components of this summary table in words.

The model show s an intercept of 11.66 which is not useful because it appears that only 6 o-rings are observed in a shuttle mission. However, the temperature shows that as the temperature increases the likelyhood of failure is decreased by .2162 for each increase in temperature.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

(c) Write out the logistic model using the point estimates of the model parameters.

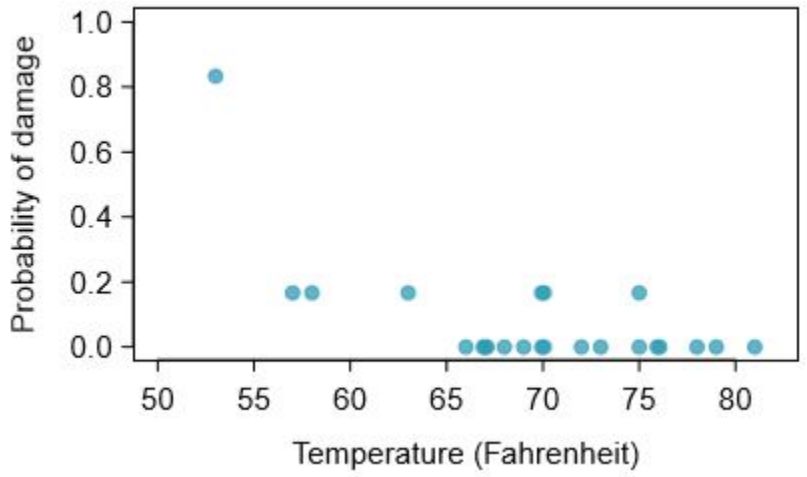
$log(\frac{\hat{p}_i}{1-\hat{p}_i}) = 11.6630 - 0.2162 * Temperature$

(d) Based on the model, do you think concerns regarding O-rings are justified? Explain.

Yes, based on the model I believe that concerns regarding o-rings are justified. Clearly, this model explains that increases in `Temperature` has a negative association with defects which makes it an important variable in reducing the probability of failures.

8.18 Challenger disaster, Part II.

Exercise 8.16 introduced us to O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into takeoff in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle. See this earlier exercise if you would like to browse the original data.



(a) The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged. Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$$\hat{p}_{57} = 0.341 \qquad \hat{p}_{59} = 0.251 \qquad \hat{p}_{61} = 0.179 \qquad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \qquad \hat{p}_{67} = 0.056 \qquad \hat{p}_{69} = 0.037 \qquad \hat{p}_{71} = 0.024$$

$\hat{p_{51}} = 11.6630 - (.2162 * 51)$

$\hat{p_{51}} = 0.6540297$

$\hat{p_{53}} = 11.6630 - (.2162 * 53)$

$\hat{p_{53}} = 0.5509228$

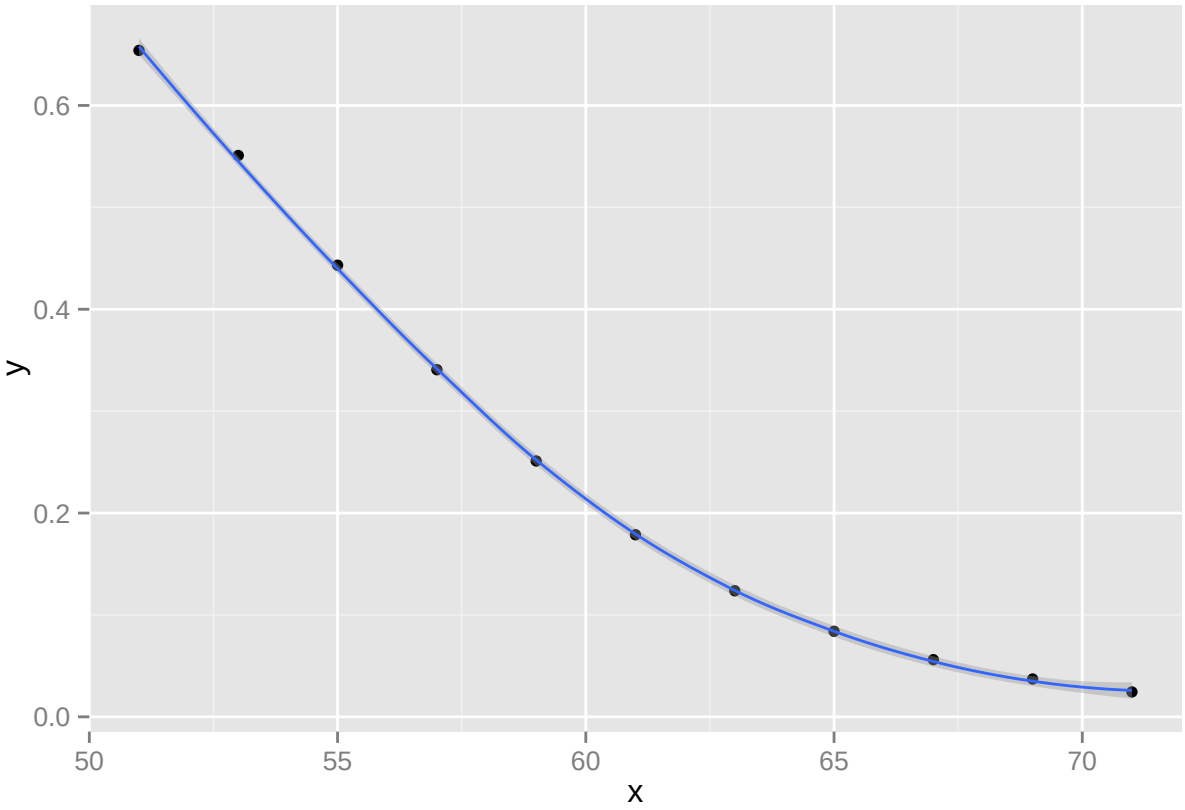
$\hat{p_{55}} = 11.6630 - (.2162 * 55)$

$\hat{p_{55}} = 0.4432456$

(b) Add the model-estimated probabilities from part (a) on the plot, then connect these dots using a smooth curve to represent the model-estimated probabilities.

```
library(ggplot2)
x <- c(51, 53, 55, 57, 59, 61, 63, 65, 67, 69, 71)
y <- data.frame(x)
df <- transform(y, y = (e ^ (11.6630 - (.2162 * y$x)) / (1 + e ^ (11.6630 - (.2162 * y$x)))))
qplot(x,y, data = df, geom = c("point", "smooth"))
```

```
## geom_smooth: method="auto" and size of largest group is <1000, so using loess. Use 'method = x' to cl
```



(c) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.

My concern about applying logistic regression in this application is that the sample size is very small and there are a number of other variables related to the o-rings that would need to be considered. It would be important to look at manufacturer and other changing variables. If the temperature was increasing at the same rate as the date of launch we might see more of an improvement in the quality production of the o-ring than just a temperature shift.